AFRL-SR-AR-TR-08-0183

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02/19/2008	FINAL PROGRESS REPORT	05/01/2007 - 11/30/2007

4. TITLE AND SUBTITLE

SPOKEN WORD RECOGNITION BY HUMANS: A SINGLE- OR A MULTI- LAYER PROCESS?

5a. CONTRACT NUMBER

FA9550-07-1-0426

5b. GRANT NUMBER

N/A

√ **5c. PROGRAM ELEMENT NUMBER**

N/A

6. AUTHOR(S)

GHITZA, ODED

5d. PROJECT NUMBER

N/A

5e. TASK NUMBER

N/A

5f. WORK UNIT NUMBER

N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SENSIMETRICS CORPORATION
48 GROVE STREET - SUITE 305
SOMERVILLE, MA 02144-2500

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
975 NORTH RANDOLPH STREET ROOM 3112
ARLINGTON, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT

[redacted] Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This 7-month-long project quantifies the role of brain rhythms in speech perception by measuring intelligibility of spoken sentences with judiciously manipulated changes in syllabic rhythm. Speech was time-compressed by a factor of three, resulting in a signal with a syllabic rate three times faster than the original and with poor intelligibility (< 50% words correct). An artificial "syllabic" rate was then introduced by segmenting the time-compressed speech signal into consecutive 40-ms intervals, each followed by a variable interval of silence. The parameters of interest were the length of the silent intervals inserted (ranging between 0-160 ms) and whether the intervals were equal in length (i.e., periodic) or not (i.e., aperiodic). The resulting performance curve is U-shaped, with best intelligibility measured at silence interval of 80 ms inserted periodically. This is also the condition in which there is a significant difference in intelligibility between periodic and aperiodic insertion (the error rate of the latter is nearly twice as high). The U-shaped performance curve may reflect the operation of cortical rhythms. Optimum intelligibility is associated with waveform-energy fluctuations in the core of the theta range of neural oscillations (3-8 Hz), which is also the core range of syllabic rate in naturally spoken utterances. Poor intelligibility may reflect the mismatch between waveform-energy fluctuations and theta rhythms in the brain.

15. SUBJECT TERMS

Syllabic rate of speech. Fast speech. Intelligibility of fast speech. Intelligibility of speech with silence insertions. Brain rhythms. Cortical oscillations. Neural oscillations. Theta range of neural oscillations.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			DR. ODED GHITZA
					19b. TELEPHONE NUMBER *(Include area code)* 617-625-0600 X239

EXECUTIVE SUMMARY

Speech is an inherently rhythmic phenomenon. Phonetic constituents are articulated in syllabic "packets," spoken in cadence and reflect energy modulations between 3-20 Hz. This rhythmic property is important for both intelligibility and naturalness. The thesis of this project is that many temporal properties of spoken language reflect not merely articulatory constraints but also higher-order principles of cortical function. In particular, we suggest that neural rhythmicity may play an important role in decoding the speech signal.

This 7-month-long project sought to quantify the role of brain rhythms in speech perception by measuring intelligibility (i.e., word-error rate) of spoken sentences with judiciously manipulated changes in syllabic rhythm. The spoken material comprised 96 semantically unpredictable sentences (SUS), each approximately 2 s long (6-8 words per sentence), spoken fluently and generated by a Text-To-Speech (TTS) speech-synthesis engine. The TTS-generated waveform was time-compressed by a factor of three, resulting in a signal with a syllabic rate three times faster than the original and with poor intelligibility (< 50% words correct). An artificial "syllabic" rate was then introduced by segmenting the time-compressed speech signal into consecutive 40-ms intervals, each followed by a variable interval of silence. The parameters of interest were the length of the silent intervals inserted (ranging between 0-160 ms) and whether the intervals were equal in length (i.e., periodic) or not (i.e., aperiodic).

The resulting performance curve (word error-rate as a function of insertion interval) is U-shaped. The highest intelligibility is associated with the condition in which the silence interval is 80 ms long and inserted periodically. This is also the condition in which there is a significant difference in intelligibility between periodic and aperiodic insertion (the error rate of the latter is nearly twice as high).

These results are surprising, and provide potential insights into how the speech signal is decoded by the brain. In our view, the U-shaped performance curve may reflect the operation of cortical rhythms. Optimum intelligibility (80 ms silence intervals inserted periodically) is associated with waveform-energy fluctuations in the core of the theta range of neural oscillations (3-8 Hz), which is also the core range of syllabic rate in naturally spoken utterances. Poor intelligibility may reflect the mismatch between waveform-energy fluctuations and theta rhythms in the brain.

1. INTRODUCTION

Speech is an inherently rhythmic phenomenon. Phonetic segments are articulated in syllabic "packages," which are spoken in cadence and reflect energy modulations between 3 and 20 Hz (Greenberg, 1999, Greenberg and Arai, 2004). The intonation contour (based on the signal's fundamental frequency) is also rhythmic (e.g. Ladd, 1996, Liberman, 1975). This rhythmic aspect of speech is important for conveying intelligibility and naturalness (for example, synthesis studies have shown that listeners prefer speech with a natural, rhythmic structure, e.g., van Santen et al., 1997). It is natural to ask whether this rhythmic property reflects something more basic that is internal to the brain.

In our view, many properties of spoken language are likely to reflect higher-order cortical processing rather than merely biomechanical and articulatory constraints. The specific patterns of articulation may themselves reflect constraints imposed by cortical function. For example, the range of time intervals (40–1000 ms) associated with different levels of linguistic abstraction (phonetic feature, segment, syllable, word, metrical foot and prosodic phrase) may reflect temporal constraints associated with neural circuits in the cerebral cortex and hippocampus. The time course of information processing in the brain is likely to constrain how fast (or slow) speech needs to be in order to be comprehensible and robust.

Neural rhythms reflect synchronous activity (both excitatory and inhibitory) in both local and widespread regions of the cortex (Buszáki, 2006). The range over which these rhythms operate (1–80 Hz) may serve as the basis for a hierarchical synchronization mechanism by which the central nervous system (CNS) processes and integrates linguistic information. It may also reflect a hierarchy of spatial scales, in which higher levels of processing depend upon information from more extensive cortical areas (von Stein and Sarnthein, 2000). For example, beta (β; 12–30 Hz) and gamma (γ; 30–80 Hz) rhythms may be involved in neural processing of phonetic segments and features, theta (θ; 3–10 Hz) rhythms in the processing of syllables and words, and delta (δ; 1–3 Hz) rhythms in processing sequences of syllables and words embedded within the metrical foot and prosodic phrase. Recent research suggests that δ frequency (< 3 Hz) oscillations may be important for certain aspects of spoken language processing (Roehm et al., 2004).

Such neural rhythms may play an important role in spoken-language comprehension (e.g. Bastiaansen and Hagoort, 2006: Brown and Hagoort (1999); Haarman et al., 2002). A variety of brain-imaging methods (e.g. PET, fMRI; e.g. Pulvermueller, 1999) enables scientists to visualize the topographic patterns of neural activation associated with linguistic processing in different regions of the cortex. The specific timing of activation across the cortex can be visualized with electromagnetic recordings (i.e. magneto-encephalography and electro-encephalography). Typically, an increase in oscillatory activity is observed in specific frequency bands, depending on the task. Of particular importance are the gamma (30-80 Hz) and the theta (3-10 Hz) rhythms (e.g. Bastiaansen et al., 2002; Bastiaansen and Hagoort, 2003; Gevins et al., 1997). Theta oscillations are most closely associated (linguistically) with the syllable (mean duration 200 ms, core range between 100 and 300 ms; Greenberg 1999), and are thought to involve some form of sensory-memory comparison process. Gamma oscillations are most closely associated with units important for diphone and other phonetic analyses.

The *precise* role of neural rhythmicity is uncertain (Buzsáki, 2006). We believe that the range of frequencies associated with such oscillatory behavior may serve as a means by which the brain integrates linguistic and other biologically important information in hierarchical fashion. As an initial hypothesis, we suggest that beta and gamma oscillations are most closely associated with linguistic processing at the phonetic-segment level, while theta oscillations are probably most closely tied to syllables and certain forms of words. Longer linguistic units, such as the metrical foot and prosodic phrase are probably associated with delta oscillations (e.g. Roehm et al., 2004).

The goal for this 7-month-long project was to demonstrate, in a quantitative manner, the importance of syllabic rhythm in spoken sentences by measuring intelligibility (i.e., word-error rate) across judiciously selected temporal-distortion conditions. Our experiments were inspired by Huggins' study on "Temporally

segmented speech" (1975). Huggins' original (i.e., unprocessed) signals were spoken passages (ca. 150 words long). These were manipulated by *inserting* silent intervals of variable length in a periodic manner (note that in insertions, no speech information is lost – in contrast to *interruptions* where a certain proportion of the speech signal is discarded). Using a "shadowing" paradigm, Huggins measured word error rate as a function of speech-time and silence-time interval durations. Deterioration in performance was hypothesized to result from the inability to bridge across the silence intervals due to the finite length of the internal memory buffer (see our proposal "Decoding Speech Using Neural Rhythmicity and Synchrony" for detail).

We offer an alternative hypothesis. In our view, the intelligibility decline is the result of a disruption in the syllabic rhythm beyond the limits of cortical neural circuitry. To quantify this hypothesis we have used a modified version of Huggins' experiment, testing the word error rate of sentences spoken fluently in two separate silence-insertion conditions: (1) *Periodic* – a speech interval followed by a silence interval, both with a prescribed fixed duration, and (2) *Aperiodic* – a speech interval of fixed duration followed by a silence interval of variable (quasi-randomly distributed) duration. A comparable error rate for the two conditions would be in line with Huggins' hypothesized role of memory-buffer limitations. In fact, we have found that there is a significant difference in intelligibility for the two conditions.

2. EXPERIMENT

2.1. Database

The database comprises 96 SUS sentences, about 2 seconds long each (6-8 words per sentence), spoken fluently. SUS stands for Semantically Unpredictable Sentences; there is no semantics, yet each individual word is associated with a meaning and the sentences follow grammatical rules.

2.2. Stimulus preparation

The original waveform was time-compressed by a factor of 3 using a pitch-synchronous, overlap and add (PSOLA) method (this is one of the features provided by the ***Praat*** software package, available online for free at http://www.fon.hum.uva.nl/praat/). Note that the resulting waveform has time-compressed formant trajectories and fricative durations but maintains the original intonation (i.e., pitch) contour. In figure 1, panel (b) shows the time-compressed version of the original, shown in panel (a).

The time-compressed waveform was the baseline for the silence-insertion. First, the baseline waveform was segmented to consecutive 40-ms long intervals. This segmentation remained fixed throughout the experiment. The silence intervals were inserted next. The parameter here is the duration of the inserted silence intervals. The conditions we used are summarized in Table 1.

Table 1. Experimental Conditions

Condition	*Speech interval, ms*	*Silence interval, ms*	*Silence/Speech*	*Speed*
x0	40	0	0	3
x20	40	20	0.5	2
x40	40	40	1	1.5
x80	40	80	2	1
x120	40	120	3	0.75
x160	40	160	4	0.6

To smooth out the abrupt transitions, every 40-ms speech interval was multiplied by a smoothing window with a rise-cosine/fall-cosine time of 1ms. A speech-spectrum-shaped noise was added to the signal after insertion; the noise intensity was adjusted to an SNR of 30 dB relative to the power of the signal prior to the insertions (i.e., condition x0). Panels (b), (c) and (d) in figure 1 show the waveforms for conditions x0, x40 and x80, respectively. See the figure caption for more details.

The conditions listed in Table 1 define the *Periodic* class of conditions. We also created an *Aperiodic* class where, per condition, the silence-time interval was of variable duration (quasi-randomly distributed), with a mean interval equal to the prescribed silence interval of Table 1, and bounded between 0.4 and 1.6 of the mean.

2.3. Subjects

All five of the subjects were young adults (graduate students at BU) with normal hearing. A prerequisite was to have all school years of education acquired in the U.S.

2.4. Instructions to subjects

Subjects performed the experiment in their home/office environment, using headphones. Each subject performed two listening sessions, "Training" and "Testing", roughly 30 minutes long each. In Training she/he listened to the original, unprocessed sentences (96 sentences). In Testing she/he listened to the same 96 sentences, processed; the 96 sentences were divided into 12 groups of 8 sentences each, six groups for Periodic (covering the list of conditions in Table 1) and six for aperiodic.

Below is an email I sent to each subject:

> Hi ___,
> I've created a web page which includes all directories/files needed for this experiment:
> http://sens.com/oded_experiment/
> 1. Download zipped file Experiment_100.1
> 2. Unzip and you a directory with 5 files.
> 3. Unzip Train and Test.
> 4. The instructions PDF file explains it all.
> 5. You may perform the experiment in your home environment.
> 6. YOU MUST LISTEN VIA HEADPHONES
> 7. Call (or email) if you have any questions. Call anytime.
> 8. Just to emphasize, you are allowed to listen to each sentence only ONCE

The instructions PDF file reads the following text:

> **Oded's intelligibility experiments – Instructions**
> The database comprises 96 SUS sentences, about 2 seconds long each, spoken fluently. SUS stands for Semantically Unpredictable Sentences; there is no semantics, yet the words are with meaning and the sentences follow grammatical rules.
>
> You are going to perform two listening sessions, "Training" and "Testing", roughly 30 minutes long each. In Training you will listen to the original, unprocessed sentences (96 sentences, stored in directory Train). In Testing you will listen to distorted (chopped) speech material (96 sentences, stored in directory Test), with order of presentation such that you first hear sentences spoken rapidly ("fast speech") – these are very hard to decode. As you proceed with sentence I.D. the effective speech rate decreases until the end of the list.
>
> You must start with Training. You may have a break between the Training session and the Testing session (each should last no longer than 30 minutes). Repeat the following for each session:
> 1. List the files in ascending order and follow the order of the List.
> 2. Open the attached text file, named text_train or text_test.
> 3. Listen to the first stimulus ONCE and type the words you have heard, even you are uncertain about.
> 4. Repeat for all stimuli.
> 5. Email the text files to me, at oded@sens.com
> 6. Thank you for participating in the experiment.

2.5. Results

2.5.1 Overall

In the Training phase, word error-rate was less than 2%. Figure 2 shows the mean performance (averaged over subjects), in terms of word error-rate, as a function of insertion interval. In the absence of insertions (condition x0), intelligibility is poor (< 50% words correct). Intelligibility is equally poor (or worse) when the insertion interval is 160 ms (condition x160). Interestingly, for insertion intervals between 20 and 120 ms, intelligibility is considerably higher. This is particularly true when the silence interval is 80 ms and inserted periodically. This is also the condition in which there is a significant difference in intelligibility between periodic and aperiodic insertion (the error rate of the latter is nearly twice as high). Two points are noteworthy. First, throughout all conditions the spectro-temporal information of the speech intervals is time-compressed by 3. Thus, the U-shape behavior is an unexpected result that is difficult to explain with conventional models. Second, the results indicate a preference for a periodic syllabic rate (at least for silence

insertion of 80 ms). Such result is also difficult to explain with conventional models. Finally, an ANOVA (analysis of variance) quantifies the significance of these trends.

2.5.2 ANOVA

Mauchly's test for sphericity revealed that assumptions of sphericity were not violated (recall that violations of this assumption can lead to invalid analysis conclusions).

The omnibus repeated-measures 2-way (2-variable) ANOVA, with significance (alpha) level of 0.05 (or 5%) showed that:

1) There is a significant main effect of insertion interval ($F(5,20)=24.163$, $p<0.0001$).
2) There is a significant main effect of periodicity (aperiodic vs periodic) ($F(1,4)=16.231$, $p<0.05$).
3) There is no significant interaction between periodicity and insertion interval ($F(5,20)=2.371$, $p>0.05$).

Post-hoc Tukey/Kramer tests showed the following significant pair-wise differences:

1) Periodic significantly different from aperiodic condition (collapsed across insertion intervals).
2) Significant differences across insertion-interval conditions (collapsed across periodicity conditions):
   a) 0 different from 20, 40, 80, 120; in addition,
   b) 20 different from 160
   c) 40 different from 160
   d) 80 different from 160
   e) 120 different from 160.

2. SUMMARY

This 7-month-long project sought to quantify the role of brain rhythms in speech perception by measuring intelligibility (i.e., word-error rate) of spoken sentences with judiciously manipulated changes in syllabic rhythm. The results are surprising and provide potential insights into how the speech signal is decoded by the brain. The U-shaped performance curve may reflect the operation of cortical rhythms. Optimum intelligibility (80 ms silence intervals inserted periodically) is associated with waveform-energy fluctuations in the core of the theta range of neural oscillations (3-8 Hz), which is also the core range of syllabic rate in naturally spoken utterances. Poor intelligibility may reflect the mismatch between waveform-energy fluctuations and theta rhythms in the brain.

In our view, endogenous cortical rhythmicity may hold the key to understanding how the brain decodes the speech signal, particularly in challenging listening conditions.

REFERENCES

Bastiaansen, M. and Hagoort, P. (2003) Event-induced theta responses as a window on the dynamics of working memory. Cortex 39: 967-992.

Bastiaansen, M. and Hagoort, P. (2006) Oscillatory neuronal dynamics during language comprehension. *Prog. Brain Res.* 159: 179-196.

Bastiaansen, M.C., Berkum, J.J. and Hagoort, P. (2002) Event-related theta power increases in the human EEG during online sentence processing. Neuroscience Letters 19: 323: 13-16.

Brown, C. and Hagoort, P. (eds.) (1999) *The Neurocognition of Language*. Oxford: Oxford University Press.

Buzsáki, G. (2006) *Rhythms of the Brain*. New York: Oxford University Press.

Gevins, A. Smith, M.E., McEvoy, L. and Yu, D. (1997) High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral Cortex* 7: 374-485.

Greenberg, S. (1999) Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29: 159-176.

Greenberg, S. and Arai, T. (2004) What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.* E87: 1059-1070.

Haarman, H.J., Cameron, J.A. and Ruchkin, D.S (2002) Neuronal synchronization meeidates on-line sentence processing: EEG coherence evidence from filler-gap constructions. *Psychophysiology* 39: 820-825.

Huggins, A.W.F. (1975) Temporally segmented speech. *Perception and Psychophysics* 18: 149-157.

Ladd, R. (1996) *Intonational Phonology*. Cambridge: Cambridge University Press.

Liberman, M. (1975) *The Intonational System of English*. Ph.D. Thesis, MIT.

Pulvermueller, F. (1999) Words in the brain's language. *Behavior and Brain Science* 22: 253-366.

Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., Haider, H. (2004) Fractionating language comprehension via frequency characteristics of the human EEG. Neuroreport 15: 409-412.

Santen, J.P.H. van, Sproat, R.W., Olive, J.P. and Hirschberg, J. (eds.) (1997) *Progress in Speech Synthesis*. New York: Springer Verlag.

Stein, A. von, and Sarnthein, J. (2000) Different frequencies for different scales of cortical integration: from local gamma to long range alpha / theta synchronization. *International Jour. Psychophysiology* 38: 301-313.

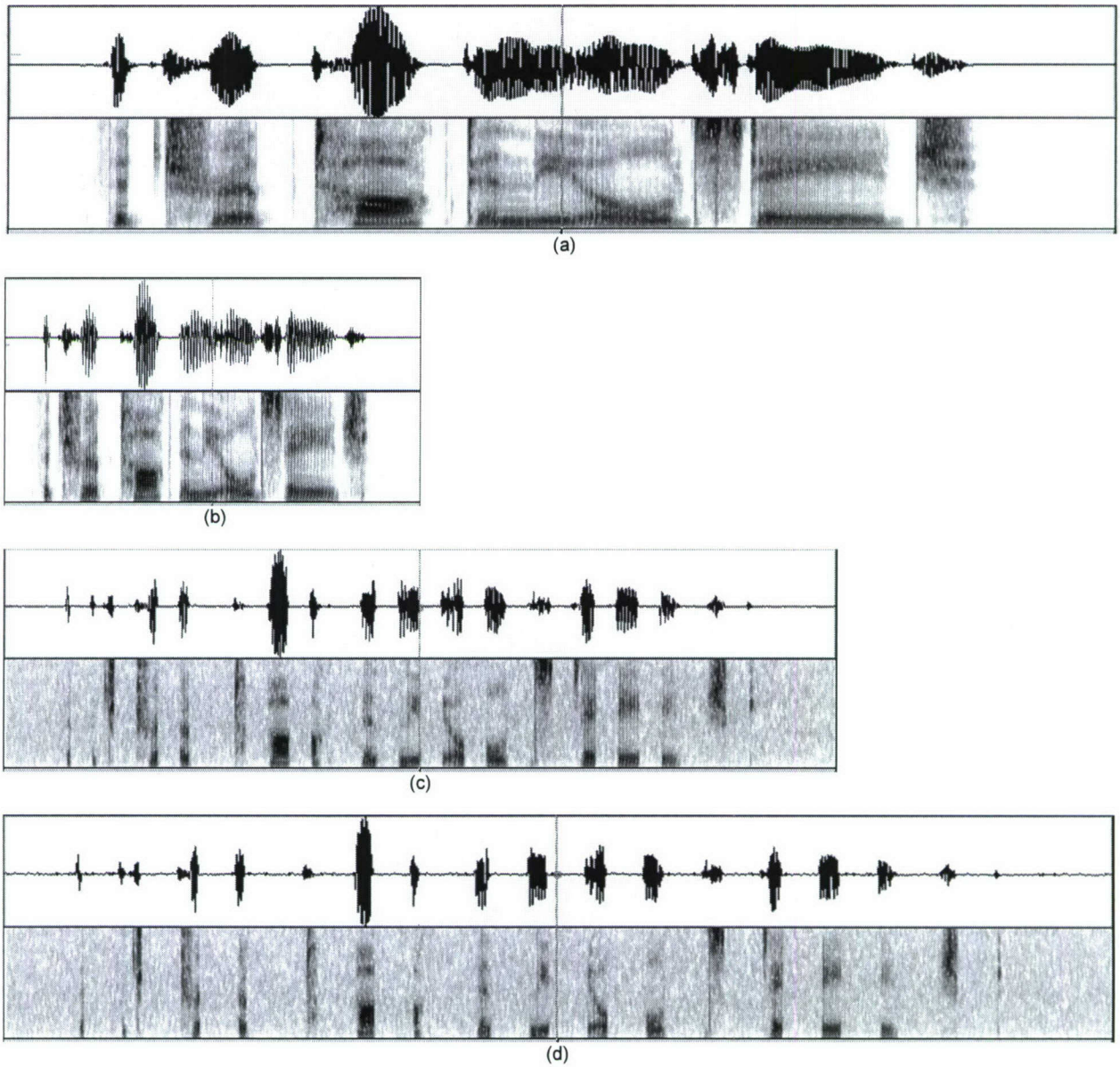

Fig. 1. (a) Waveform (top) and wideband spectrogram (bottom) of the sentence "the trip talked in the old stage". The waveform duration is 2.4 seconds, the upper frequency of the spectrogram 5000Hz; (b) Same as (a), time-compressed by a factor of 3; (c) Consecutive 40ms long speech intervals of (b), with 40ms long silence insertions. Note that the duration of the processed speech waveform is 2/3 the duration of the original (i.e. time-compressed by a factor of 1.5 re original); (d) Same as (c) with 80ms long silence intervals. The duration of the waveform is same as the original waveform duration (i.e. no time compression re original). Note that the speech intervals are identical to those in (c).

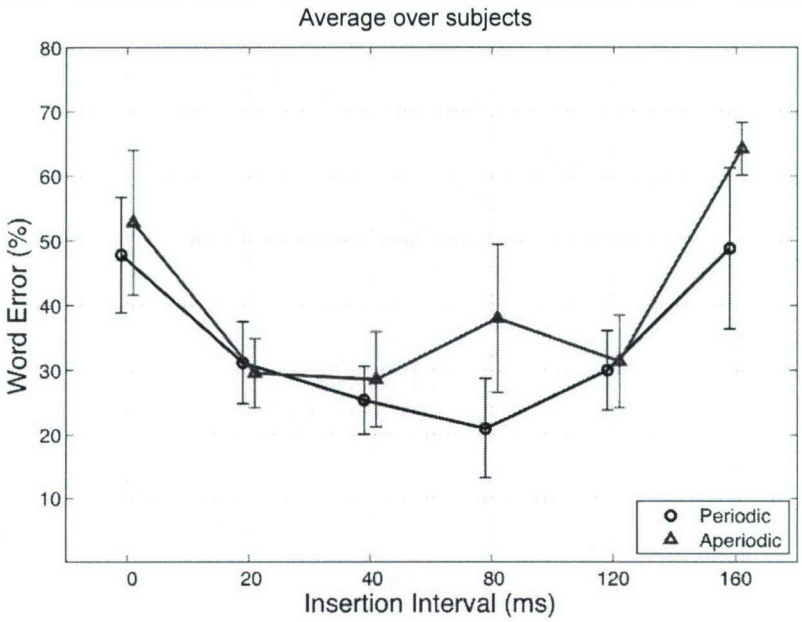

Fig. 2. Intelligibility of fast speech with silence insertions. Word error-rate is plotted as a function of the silence-interval duration. Speech is time-compressed by a factor of 3. Speech segments are consecutive 40ms long intervals, and are kept the same for all conditions.

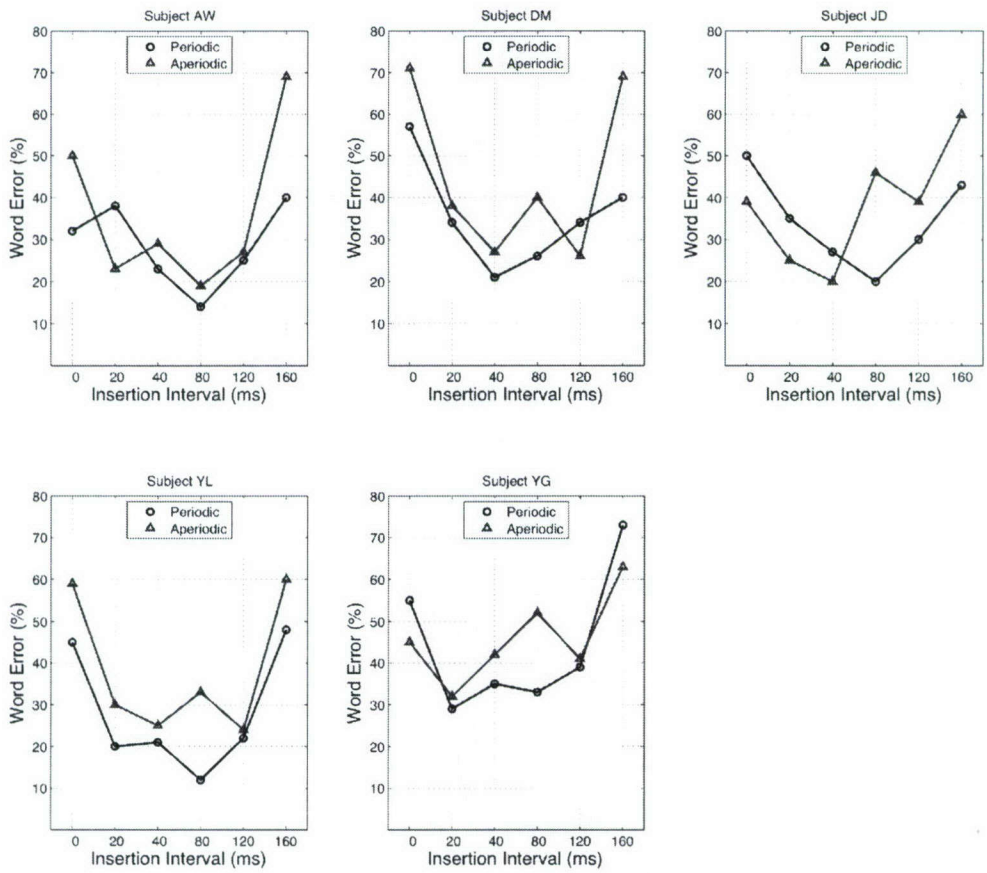

Fig. 3. Same as in Fig. 2, per subject.